МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Прогнозирование конечных свойств новых материалов

(композиционных материалов)

Слушатель Горшков Андрей Вячеславович

Москва, 2023

СОДЕРЖАНИЕ

с.

ВВЕДЕНИЕ............................................................................................................... 3

АТТЕСТАЦИОННОЕ ЗАДАНИЕ.......................................................................... 4

1 АНАЛИТИЧЕСКАЯ ЧАСТЬ................................................................................ 5

1.1 Предварительный анализ датасета.....................................................................5

1.2 Последовательность операций (пайплайн) машинного бучения................... 6

1.3 Методы**,** используемые для выполнения операций пайплайнов................... 9

1.4 Фильтрация полного датасета от помех и шума……………….....................13

2 ПРЕДОБРАБОТКА ДАННЫХ............................................................................16

2.1 Визуализация данных....................................................................................... 16

2.2 Коэффициенты корреляции и аналитическийкоэффициент детерминации для МНК……………………………………………………………………….......21

3 РЕЗУЛЬТАТЫ МАШИННОГО ОБУЧЕНИЯ МОДЕЛЕЙ.............................. 27

3.1 Результаты обучения моделей на данных базового датасета........................27

3.2 Результаты обучения моделей на данных полного датасета......................... 29

3.3 Результаты обучения моделей на данных очищенного датасета.................. 31

4 СОЗДАНИЕ УДАЛЕННОГО РЕПОЗИТОРИЯ................................................ 34

ЗАКЛЮЧЕНИЕ .......................................................................................................35

БИБЛИОГРАФИЧЕСКИЙ СПИСОК................................................................... 37

**ВВЕДЕНИЕ**

Композиционные материалы – это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т.е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита – железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

**Актуальность**: Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

**АТТЕСТАЦИОННОЕ ЗАДАНИЕ**

1. Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER

<https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing>

1. Обучить алгоритм машинного обучения, который будет определять значения:

* Модуль упругости при растяжении, ГПа
* Прочность при растяжении, МПа

1. Написать нейронную сеть, которая будет рекомендовать:

* Соотношение матрица-наполнитель

1. Написать приложение, которое будет выдавать прогноз, полученный в задании 2 или 3 (один или два прогноза, на выбор учащегося)
2. Создать профиль на github.com
3. Сделать commit приложения на github.com
4. Сделать commit на веб-хостинг (По желанию учащегося)
5. Написать пояснительную записку к проекту, которая включает блок-схему и описание процесса подготовки, обучения моделей и инструкцию по установке и запуску приложения.

**1 АНАЛИТИЧЕСКАЯ ЧАСТЬ**

**1.1 Предварительный анализ датасета**

После объединения (по индексу) отдельных частей датасета итоговый датасет содержит 1023 строки и 13 столбцов – см. рисунок 1.

В результате визуального анализа выявлено, что первые 23 строки датасета содержат преимущественно целые и рациональные числа. При этом рациональные числа являются средними значениями первых 23 строк соответствующих столбцов датасета (за исключением показателя «Соотношение матрица-наполнитель»), что позволяет сделать предположение о том, что в этих строках имелись пропуски данных, которые были заполнены средними значениями первых 23 строк соответствующих столбцов.

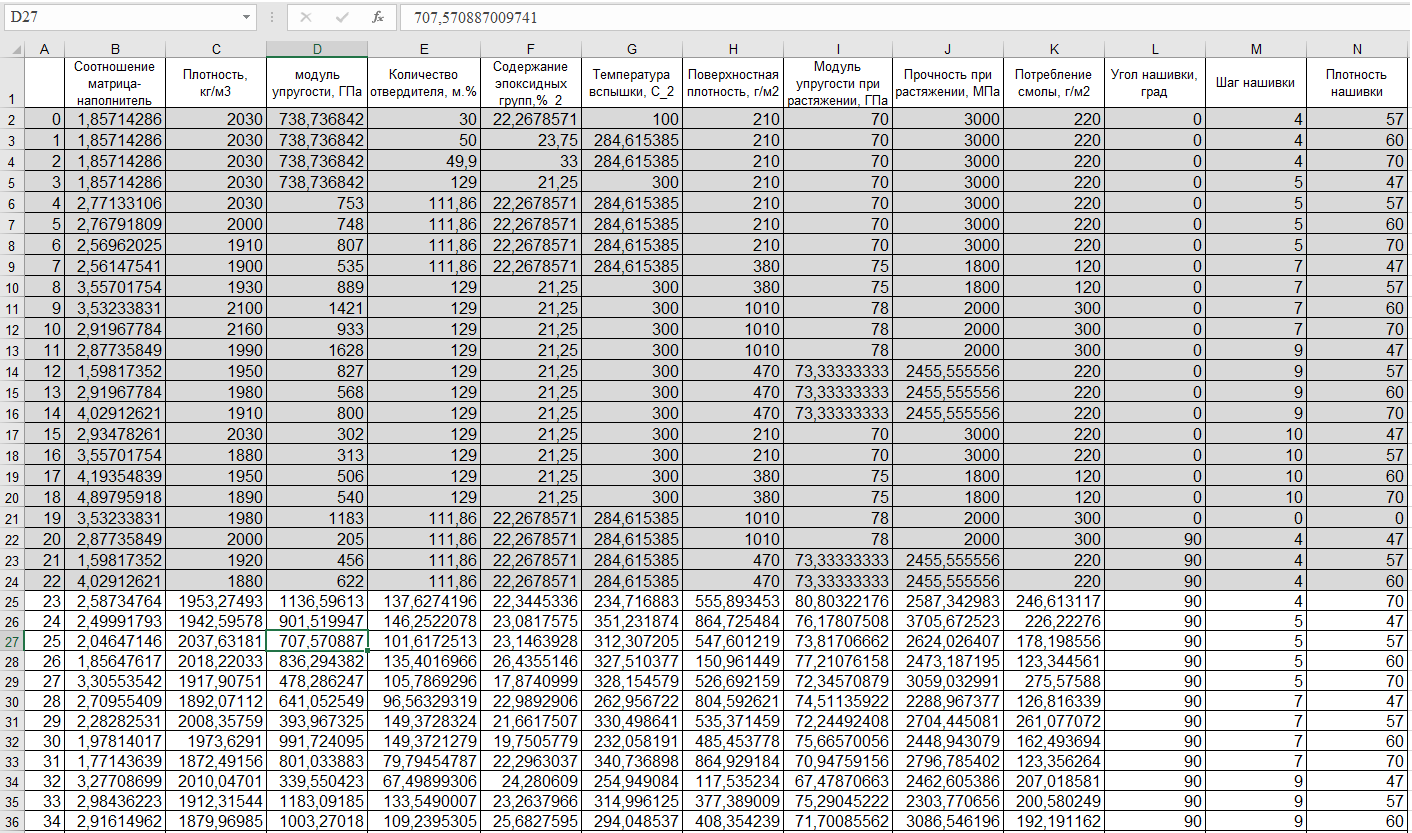


Рисунок 1 – Фрагмент датасета

Остальные 1000 строк датасета содержат 14-значные и 15-значные иррациональные числа без пропусков данных. Установлено, что средние значения и среднеквадратичные отклонения (СКО) значений столбцов этой части датасета с высокой точностью совпадают со средними значениями и СКО соответствующих столбцов, как первых 23 строк датасета, так и полного датасета. С высокой степенью уверенности можно утверждать, что 14-значные и 15-значные иррациональные числа в датасете являются результатом вычислений, а не результатом прямых или косвенных измерений (маловероятно, что средства измерений физических характеристик композитов позволяют проводить измерения с такой высокой точностью).

На основании выявленной информации можно предположить, что первые 23 строки датасета являются планом многофакторного измерительного или численного эксперимента, пропуски данных в котором были заполнены средними значениями первых 23 строк соответствующих столбцов. Следовательно, данные в этих строках являются достоверными (с некоторой точностью) значениями. Для определенности та часть датасета, которая состоит из первых 23 строк, в данной работе называется базовым датасетом.

Также можно предположить, что последние 1000 строк датасета представляют собой совокупность двух видов значений данных.

Первый вид значений – значения в промежуточных точках базового датасета, вычисленные по аппроксимациям зависимостей переменных базового датасета.

Второй вид значений – гауссовский шум, сгенерированный по средним значениям и СКО данных в соответствующих столбцах базового датасета.

Возможный третий вид значений – суперпозиция первого и второго вида значений в данной работе не рассматривается.

В итоге проведенного предварительного анализа принято следующее решение – построить математические модели целевых переменных на данных каждого из трех следующих вариантов набора данных:

1) базовый датасет (первые 23 строки);

2) полный датасет (1023 строки);

3) очищенный датасет, полученный путем фильтрации полного датасета от помех и шума (*N* строк: *N* < 1023).

**1.2 Последовательность операций (пайплайн) машинного обучения**

**1.2.1 Пайплайн машинного обучения на данных базового датасета**

Принятый в данной работе пайплайн машинного обучения на данных базового датасета состоит из следующих операций – см. рисунок 2:

1) разведочный анализ данных;

2) отбор значимых признаков;

3) обучение различных моделей на всех данных (23 строки) базового датасета;

4) вычисление коэффициента детерминации и других метрик моделей, вычисление доверительного интервала для расчетных значений лучшей модели.

Особенностью данного пайплайна является отсутствие таких процедур, как удаление выбросов и разделение базового датасета на обучающую и тестовую выборки. Основанием для этого является малое число объектов (точек) базового датасета.



Рисунок 2 – Пайплайн обработки базового датасета

**1.2.2 Пайплайн машинного обучения на данных полного датасета**

Принятый в данной работе пайплайн машинного обучения на данных полного датасета состоит из следующих операций – см. рисунок 3:

1) разведочный анализ данных;

2) удаление выбросов;

3) нормализация данных (приведение к нормальному закону распределения);

4) отбор значимых признаков;

5) разбиение на обучающую (train) и тестовую (test) выборки;

6) масштабирование признаков;

7) обучение различных моделей на всех данных полного датасета;

8) вычисление коэффициента детерминации и других метрик моделей, вычисление доверительного интервала для прогноза значений лучшей модели.

Для исключения путаницы в терминологии в данной работе под нормализацией данных понимается процедура приведения распределения данных к закону распределения, близкому к нормальному.



Рисунок 3 – Пайплайн обработки полного датасета

**1.2.3 Пайплайн машинного обучения на данных очищенного датасета**

Принятый в данной работе пайплайн машинного обучения на данных очищенного датасета состоит из следующих операций ­ – см. рисунок 4:

1) разведочный анализ данных

2) очистка (фильтрация) полного датасета от помех и шума;

3) нормализация данных (приведение к нормальному закону распределения);

4) отбор значимых признаков;

5) разбиение на обучающую (train) и тестовую (test) выборки;

6) масштабирование признаков;

7) обучение различных моделей на всех данных полного датасета;

8) вычисление коэффициента детерминации и других метрик моделей, вычисление доверительного интервала для прогноза значений лучшей модели.

Данный пайплайн отличается от пайплайна обучения на данных полного датасета только заменой процедуры удаления выбросов процедурой фильтрации датасета от шума и помех.



Рисунок 4 – Пайплайн обработки очищенного датасета

**1.3 Методы, используемые для выполнения операций пайплайнов**

**1.3.1 Разведочный анализ данных**

При разведочном анализе данных выполнялись следующие процедуры – визуализация данных и вычисление статистических характеристик датасета.

В рамках визуализации данных проводилось построение гистограмм распределения каждой из переменной, попарных графиков рассеяния точек, диаграмм boxplot («ящиков с усами»).

Кроме того, вычислены основные статистические характеристики всех целевых переменных и признаков датасета:

- средние значения;

- медианные значения;

- среднеквадратичные отклонения;

- другие статистические параметры.

**1.3.2 Удаление выбросов в данных**

Для поиска выбросов использовались следующие методы:

- метод трех среднеквадратичных отклонений (СКО) s – обнаружение данных, значения которых отклоняются от средних значений более, чем на 3s;

- метод межквартильного размаха **–** обнаружение данных, значения которых расположены за пределами третьего (Q3) и первого (Q1) квартилями;

- метод Isolation Forest – алгоритм для обнаружения [аномалий](https://translated.turbopages.org/proxy_u/en-ru.en.02c2b4af-643d489e-6d722fb2-74722d776562/https/en.wikipedia.org/wiki/Anomaly_detection) в данных (выбросов), использующий решающие [бинарные деревья](https://translated.turbopages.org/proxy_u/en-ru.en.02c2b4af-643d489e-6d722fb2-74722d776562/https/en.wikipedia.org/wiki/Binary_tree) [1]. Идея алгоритма заключается в том, что при разделении пространства данных линиями, [ортогональными](https://translated.turbopages.org/proxy_u/en-ru.en.02c2b4af-643d489e-6d722fb2-74722d776562/https/en.wikipedia.org/wiki/Orthogonality) началу координат, те точки данных, для изоляции которых от датасета требуется меньше разделений, с высокой вероятностью являются аномалиями данных (особенно точки, являющиеся листьями).

**1.3.3 Нормализация данных**

В полном датасете имеются нулевые значения некоторых признаков. По этой причине для нормализации данных использовалось преобразование Йео – Джонсона [2], позволяющее трансформировать нулевые и отрицательные значения.

**1.3.4 Отбор значимых признаков**

Для отбора значимых признаков использовались следующие операции:

- вычисление коэффициентов корреляции Пирсона и Спирмена;

- вычисление статистических показателей значимости признаков – взаимной информации MI (mutual information) и F-теста;

- анализ мультиколлинеарности признаков путем вычисления коэффициента увеличения дисперсии VIF (variance inflation factor);

- пошаговый отбор признаков (stepwise regression).

**1.3.5 Разбиение на обучающую и тестовую выборки**

Разбиение данных на обучающую (train) и тестовую (test) выборки проводилось в соотношении 70/30.

Для предотвращения пере- и недообучения моделей (с целью обеспечения их приемлемой обобщающей способности) проводился контроль подобия обучающей и тестовой выборок. Контроль проводился путем сравнения значений аналитических коэффициентов детерминации для обучающей и тестовой выборок.

Аналитический коэффициент детерминации является коэффициентом детерминации для линейной модели, построенной методом наименьших квадратов (МНК), и вычисляется по значениям коэффициентов корреляции выборок следующим образом [3]

где – определитель матрицы парных коэффициентов корреляции между целевой переменной и признаками;

– определитель матрицы парных коэффициентов межфакторной корреляции (между признаками).

Предполагается, что если обучающая и тестовая выборки подобны (то есть имеют сравнимые коэффициенты корреляции и, следовательно, сравнимые коэффициенты детерминации), то это будет способствовать предотвращению пере- и недообучения моделей.

В данной работе при существенном – на порядок и более – различии аналитических коэффициентов детерминации обучающей и тестовой выборок данное разбиение отбрасывалось и для обучения моделей использовалось другое разбиение, при котором аналитические коэффициенты детерминации для обучающей и тестовой выборок были сравнимы.

Необходимо отметить еще одно возможное назначение аналитического коэффициента детерминации для МНК – его использование для оценки принципиальной возможности построения нетривиальной (не константной) модели для данного датасета. Предполагается, что если вычисленное по выражению (1) значение аналитического коэффициента детерминации для обучающей выборки очень мало ( 0), то построить не константную, то есть пригодную для практического применения математическую модель на этих данных практически невозможно или маловероятно для любого алгоритма.

**1.3.6 Масштабирование признаков**

Масштабирование признаков применялось для улучшения сходимости алгоритмов обучения различных моделей. В качестве методов масштабирования использовались:

- стандартизация (центрирование данных относительно их среднего значения и приведение к их среднеквадратичному отклонению);

- приведение к межквартильному диапазону

где и – первый и третий квартиль распределения данных, соответственно.

**1.3.7 Машинное обучение моделей**

Машинное обучение математических моделей в данной работе выполнено на языке программирования Python 3.

**1.3.7.1 Целевые переменные «Модуль упругости при растяжении» и «Прочность при растяжении»**

Для построения математических моделей целевых переменных «Модуль упругости при растяжении» и «Прочность при растяжении» применялись следующие алгоритмы из библиотеки scikit-learn языка программирования Python 3.

а) Параметрические алгоритмы регрессии:

- LinearRegression;

- ElasticNet.

б) Ансамблевые алгоритмы:

- GradientBoostingRegressor;

- RandomForestRegressor.

в) Непараметрические алгоритмы регрессии:

- KneighborsRegressor;

- TheilSenRegressor [4].

Для построения моделей с наилучшими характеристиками проводился поиск оптимальных значений их гиперпараметров по сетке с перекрестной проверкой с количеством блоков, равным 10.

**1.3.7.2 Целевая переменная «Соотношение матрица – наполнитель»**

Для построения математических моделей целевой переменной «Соотношение матрица – наполнитель» рассматривались полносвязные нейронные сети с двумя скрытыми слоями и одним выходным слоем.

**1.3.8 Вычисление метрик обученных моделей и доверительных интервалов для прогноза значений целевых переменных**

Для анализа точности и обобщающей способности обученных моделей вычислялись метрики , MAE, RMSE для обучающих и тестовых выборок. Для наглядности метрики обученных моделей отражены в сводных таблицах метрик разных моделей.

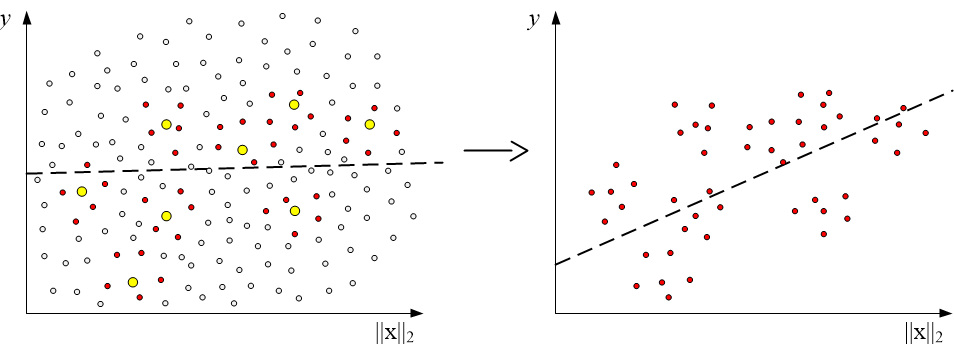
Согласно вычисленным статистическим характеристикам полного датасета все целевые переменные и признаки имеют сравнительно большую дисперсию. Следовательно, доверительные интервалы целевых переменных также будут иметь относительно большие значения, вследствие прямой пропорциональности ширины доверительного интервала от дисперсии значений переменных. Учитывая это обстоятельство принято решение рассматривать в данной работе не только точечные прогнозируемые значения целевых переменных, но и их доверительные интервалы с уровнем значимости 0,05. Разумеется, что представление результата прогноза в виде доверительного интервала позволит достоверно, то есть с определенной точностью, прогнозировать целевые переменные по обученным моделям.

**1.4 Фильтрация полного датасета от помех и шума**

Данная процедура имеет первостепенное значение для решения поставленной задачи – создания нетривиальных (не константных), то есть пригодных для практического применения математических моделей целевых переменных на основе предоставленного датасета.

В основе идеи фильтра лежит предположение, что первые 23 строки полного датасета являются планом многофакторного эксперимента. Следовательно, данные в этих строках являются достоверными, разумеется, с некоторой точностью, которая определяется некоторым доверительным интервалом. Тогда те значения данных, которые находятся в этом доверительном интервале, с высокой вероятностью также являются достоверными значениями. И наоборот, те данные, которые расположены вне доверительного интервала, с высокой вероятностью содержат значительные составляющие помех и (или) шумов и потому должны быть исключены из процесса обучения моделей.

Таким образом, объекты базового датасета можно рассматривать как центроиды, которые позволяют отобрать достоверные значения, например, методом «ближайших соседей» – см. рисунок 5.



а) полный датасет; б) очищенный датасет (*knn* = 5)

Рисунок 5 – Очистка датасета от помех и шума с помощью фильтра, ядром которого являются объекты (выделены желтым цветом) базового датасета.

Примечание. Пунктиром показаны линии регрессии

При этом предполагаем, что центроиды и «ближайшие соседи» имеют одинаковую значимость (вес), так как нет оснований полагать, что данные базового датасета являются более достоверными, чем отобранные в результате фильтрации объекты. Следовательно, для дальнейшего машинного обучения моделей центроиды и «ближайшие соседи» являются равнозначными данными.

Для улучшения процедуры фильтрации можно использовать итеративный алгоритм фильтра – сначала отбираются «ближайшие соседи» для объектов базового датасета, обучается модель регрессии, затем отбираются «ближайшие соседи» уже в некотором доверительном интервале для всей кривой регрессии, снова обучается модель и т. д., пока полученная модель регрессии не станет устойчивой. В данной работе рассматривается только первая итерация алгоритма фильтра – отбираются «ближайшие соседи» только для объектов базового датасета.

Необходимо отметить, что если все последние 1000 строк полного датасета сгенерированы случайным образом, то эту процедуру следует рассматривать не как фильтрацию датасета, а как размножение (oversampling) достоверных данных – объектов базового датасета.

**2 ПРЕДОБРАБОТКА ДАННЫХ**

Программный код и результаты предобработки данных приведены в следующих файлах, расположенных на странице автора в репозитории GitHub – см. 4.2:

first\_23.ipynb

y\_1\_full.ipynb

y\_1\_filtered.ipynb

y\_2\_full.ipynb

y\_2\_filtered.ipynb

y\_3\_full\_NN.ipynb

y\_3\_filtered\_NN.ipynb

y\_3\_filtered\_Regressor.ipynb

В программном коде и пояснительной записке приняты следующие обозначения целевых переменных.

y\_1 – 'Модуль упругости при растяжении, ГПа'

y\_2 – 'Прочность при растяжении, МПа'

y\_3 – 'Соотношение матрица-наполнитель'

**2.1 Визуализация данных**

Визуализация данных проводилась с использованием библиотек matplotlib и seaborn языка программирования Python 3.

На рисунках 6, 7, 9, 10 приведены точечные диаграммы и гистограммы распределений целевых переменных и признаков для базового, полного и очищенных датасетов.

На рисунке 8 приведен график boxplot для целевых переменных и признаков полного датасета после удаления выбросов методом Isolation Forest.

Визуальный анализ приведенных на рисунке 7 распределений позволяет сделать вывод, что полный датасет содержит большое количество данных, являющихся шумом или помехами, в результате чего корреляция между целевыми переменными и признаками практически отсутствует.

**2.1.1 Визуализация данных базового датасета**

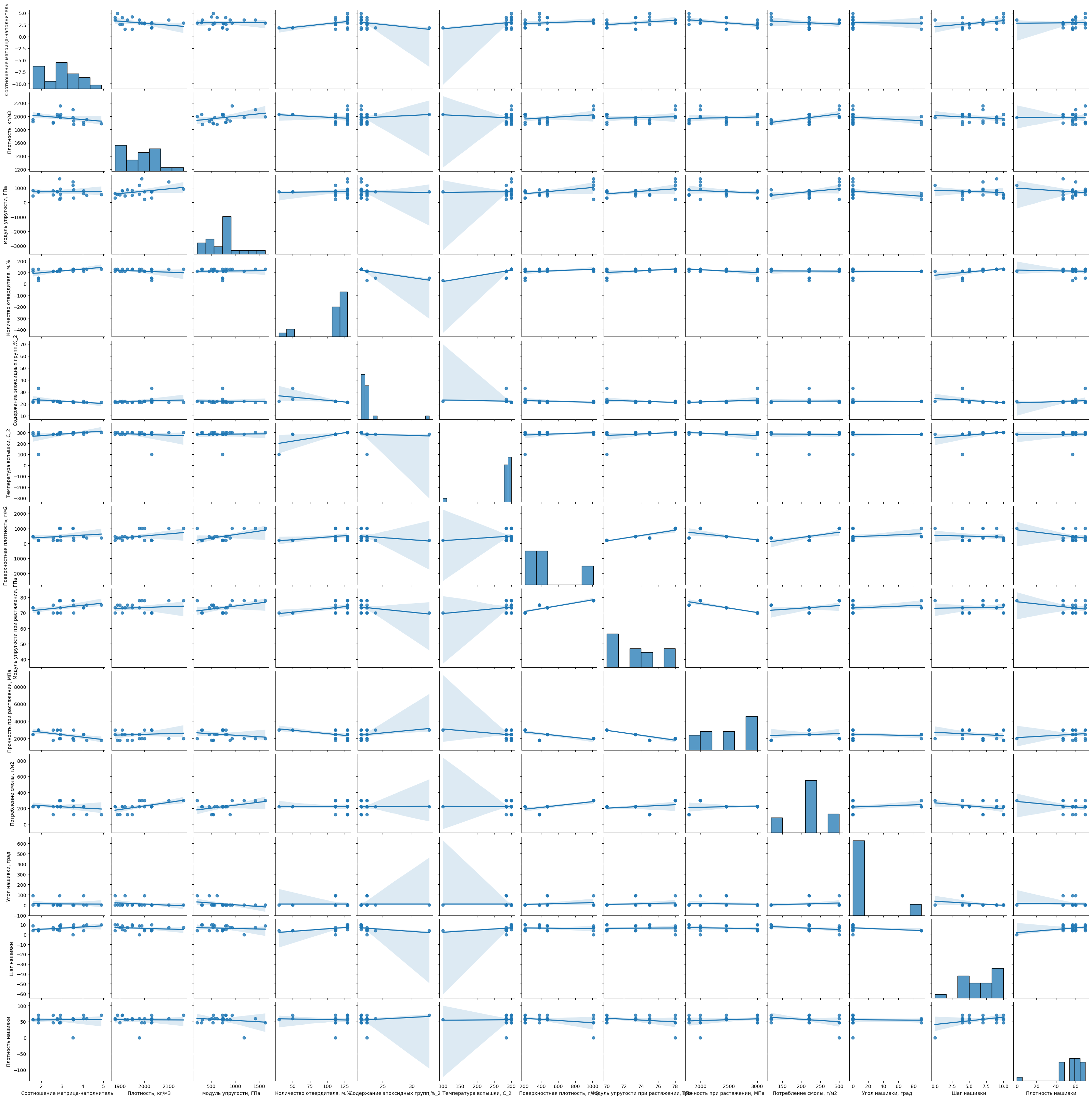


Рисунок 6 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков базового датасета

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

**2.1.2 Визуализация данных полного датасета**

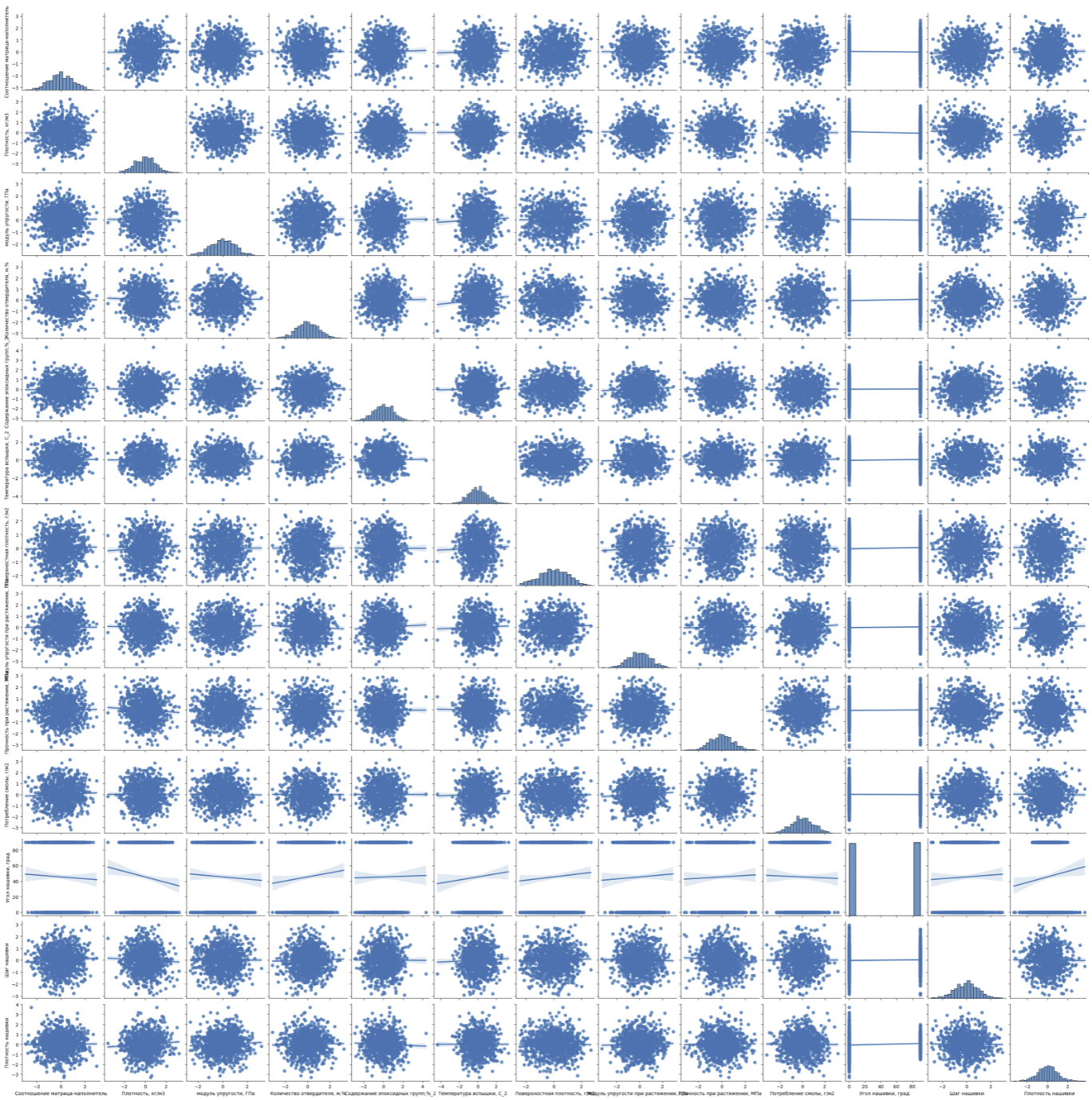
****

Рисунок 7 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков полного датасета (после удаления выбросов и преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

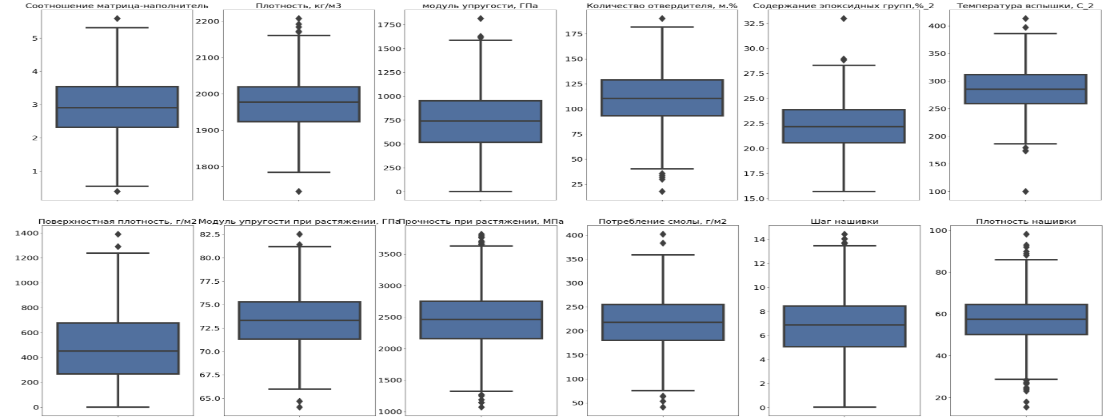
****

Рисунок 8 – Графики boxplot для целевых переменных и признаков полного датасета (после удаления выбросов методом Isolation Forest)

**2.1.3 Визуализация данных очищенных датасетов**

а) Число «ближайших соседей» *knn* = 3

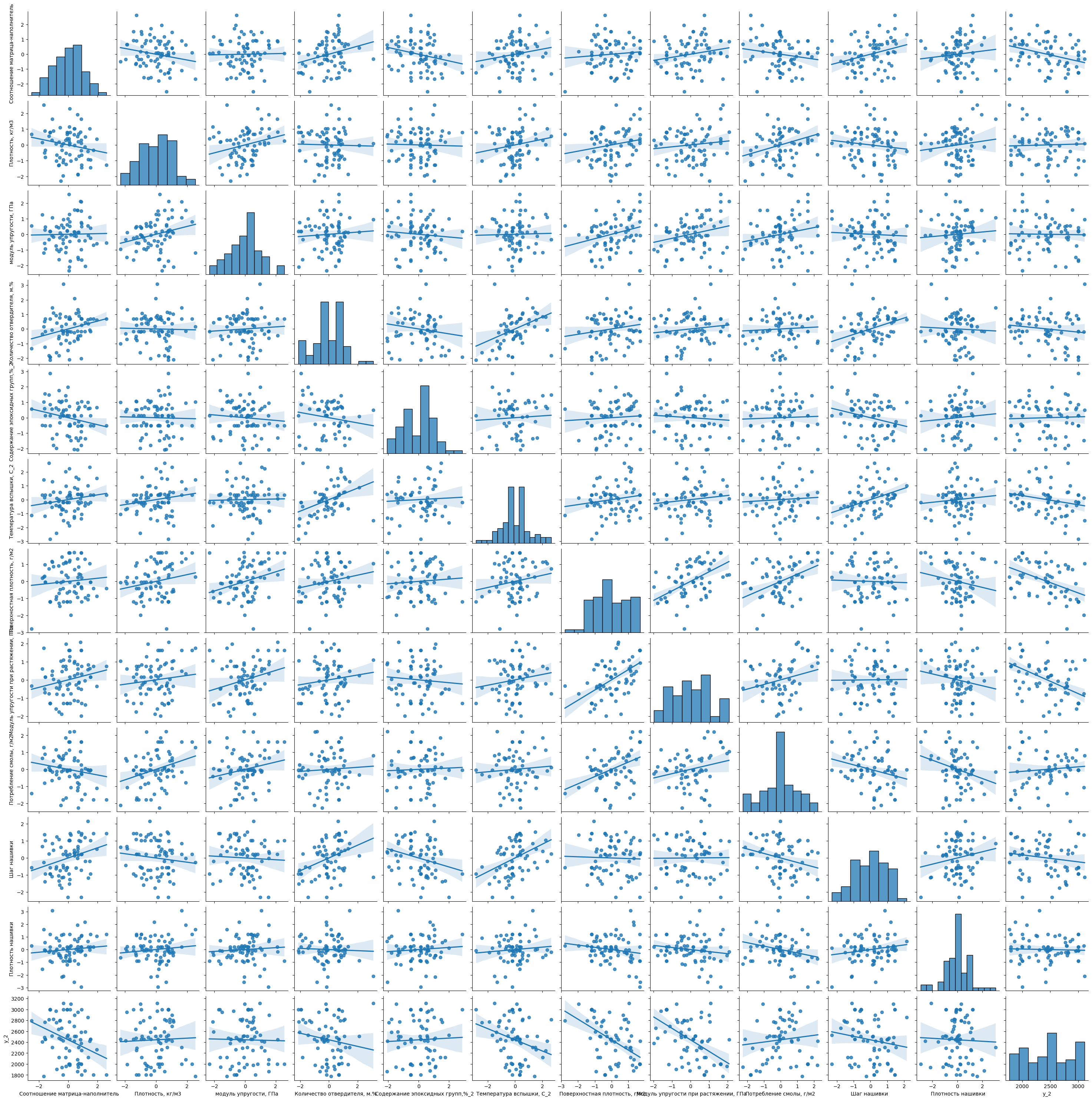


Рисунок 9 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков очищенного датасета (*knn* = 3; после преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

б) Число «ближайших соседей» *knn* = 5

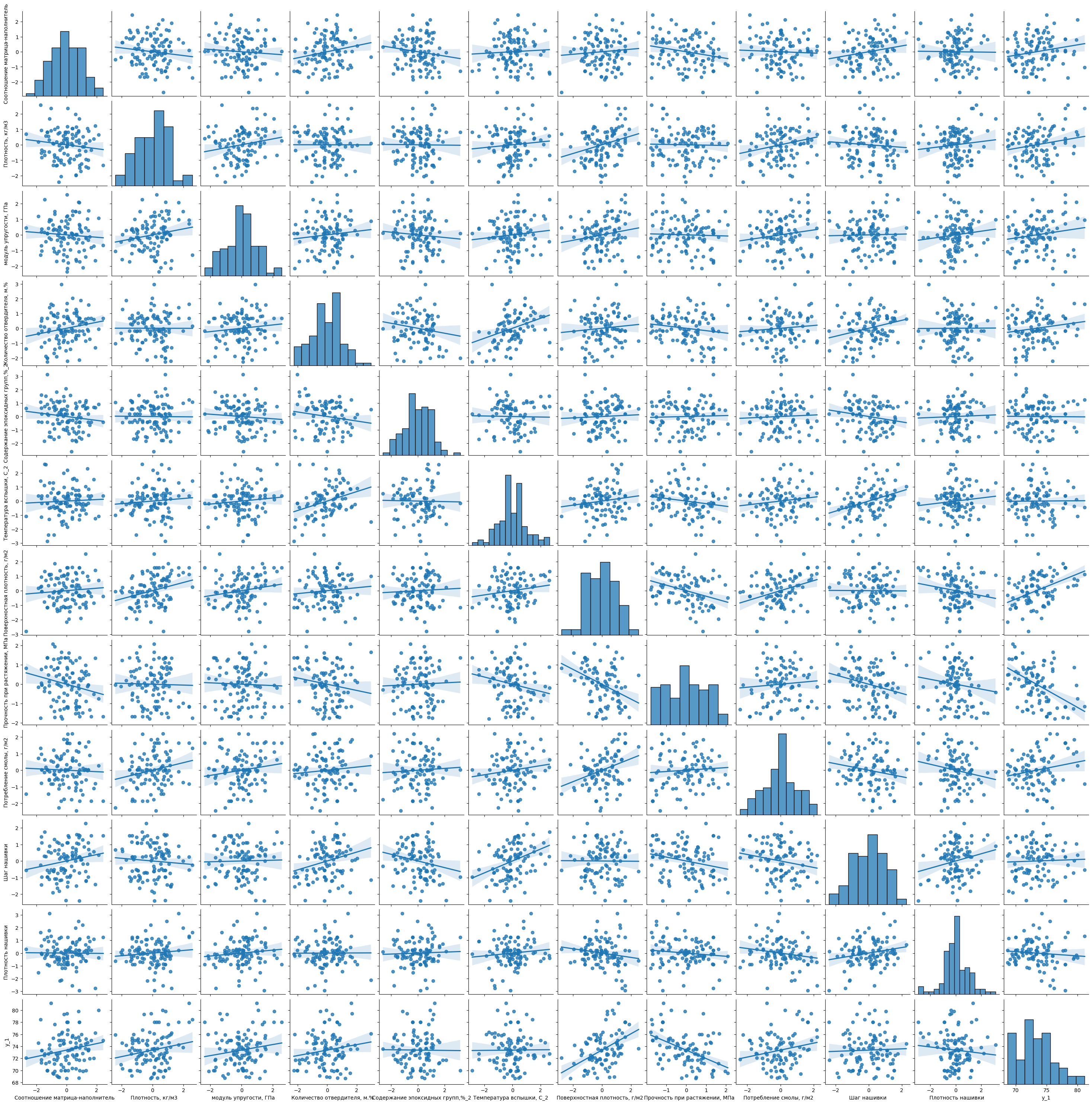


Рисунок 10 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков очищенного датасета (*knn* = 5; после преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

**2.2 Коэффициенты корреляции и аналитический коэффициент детерминации для МНК**

На рисунках 11 – 18 приведены матрицы попарных коэффициентов корреляции Пирсона для целевых переменных и признаков базового, полного и очищенного датасета.

В таблицах 1 – 3 приведены расчетные значения аналитических коэффициентов детерминации *R*2 для целевых переменных базового, полного и очищенных датасетов. Аналитические коэффициенты детерминации рассчитаны как для целых датасетов (до их разбиения на обучающую и тестовые выборки), так и для обучающих и тестовых выборок.

Согласно приведенным в таблице 1 данным для базового датасета значения аналитического коэффициента детерминации *R*2 = 1 для целевых переменных 'Модуль упругости при растяжении, ГПа' и 'Прочность при растяжении, МПа'. Следовательно, данные переменные имеют линейные функциональные зависимости с некоторыми признаками. Целевая переменная 'Соотношение матрица-наполнитель' имеет значение аналитического коэффициента детерминации *R*2 = 0,393, что позволяет сделать вывод о принципиальной возможности построения нетривиальной (не константной) модели для данной целевой переменной по объектам базового датасета.

Согласно приведенным в таблице 2 данным для полного датасета все целевые переменные имеют практически нулевые значения аналитического коэффициента детерминации *R*2, что позволяет сделать вывод о принципиальной невозможности или малой вероятности построения нетривиальных (не константных) математических моделей для целевых переменных по объектам полного датасета.

Согласно приведенным в таблице 3 данным для очищенных датасетов (при *knn* = 3 и *knn* = 5) целевые переменные имеют значения аналитического коэффициента детерминации *R*2 в диапазоне от 0,242 до 0,812, что позволяет сделать вывод о принципиальной возможности построения нетривиальных (не константных) моделей для целевых переменных по объектам очищенных датасетов.

**2.2.1 Коэффициенты корреляции и детерминации для базового датасета**

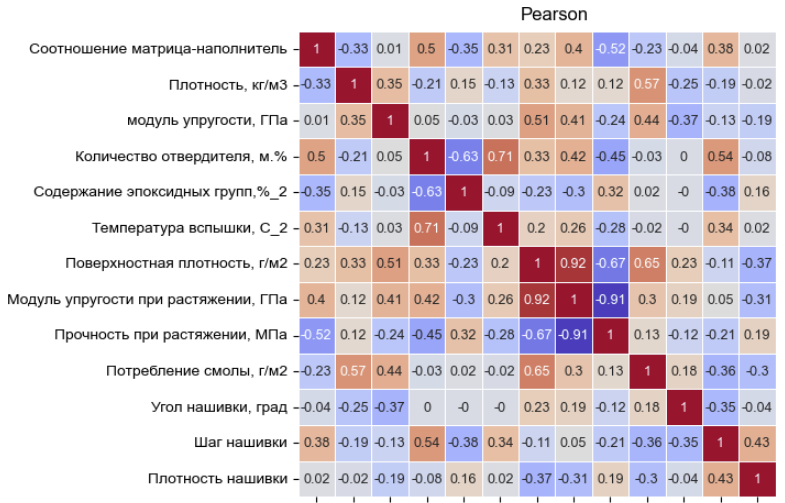
****

Рисунок 11 – Матрица корреляции признаков базового датасета

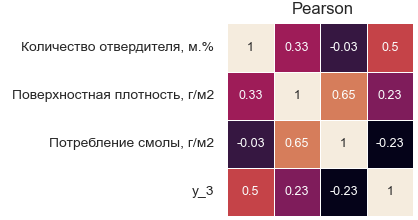
****

Рисунок 12 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (базовый датасет)

Таблица 1 – Аналитические коэффициенты детерминации для целевых переменных (базовый датасет)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Аналитический коэффициент детерминации для МНК *R*2 | | |
| до разбиения на train-  и test-выборки | train-выборка | test-выборка |
| 'Модуль упругости при растяжении' | 1 | разбиение не проводилось | разбиение не проводилось |
| 'Прочность при растяжении' | 1 | разбиение не проводилось | разбиение не проводилось |
| 'Соотношение матрица - наполнитель' | 0,393 | разбиение не проводилось | разбиение не проводилось |

**2.2.2 Коэффициенты корреляции и детерминации для полного датасета**

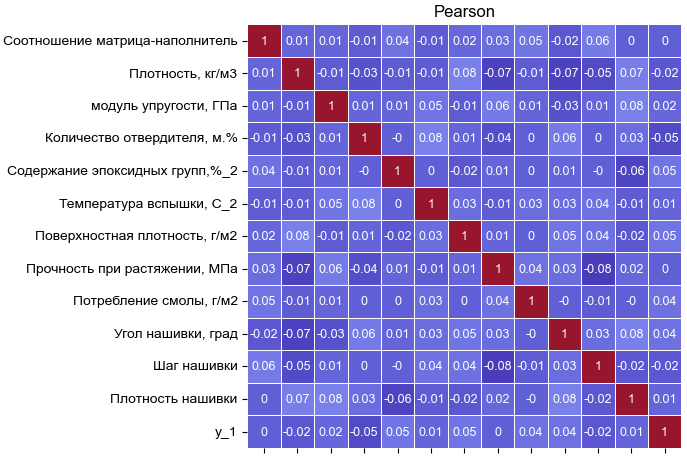
****

Рисунок 13 – Матрица корреляции признаков полного датасета

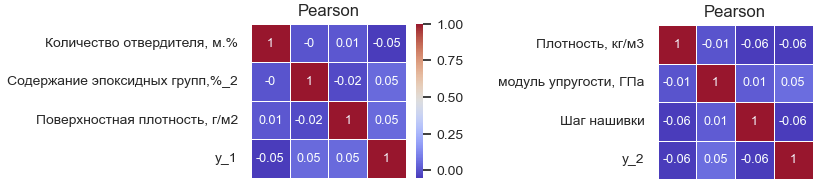
****

Рисунок 14 – Матрицы корреляции отобранных признаков для переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' (полный датасет)

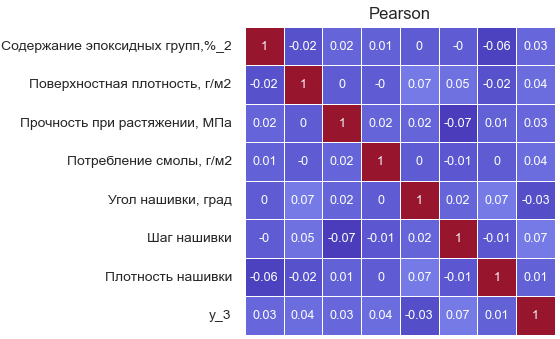


Рисунок 15 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (полный датасет)

Таблица 2 – Аналитические коэффициенты детерминации для целевых переменных (полный датасет)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Аналитический коэффициент детерминации для МНК *R*2 | | |
| до разбиения на train-  и test-выборки | train-выборка | test-выборка |
| 'Модуль упругости при растяжении' | 0,012 | 0,008 | 0,005 |
| 'Прочность при растяжении' | 0,016 | 0,009 | 0,017 |
| 'Соотношение матрица - наполнитель' | 0,012 | 0,008 | 0,057 |

**2.2.3 Коэффициенты корреляции и детерминации для очищенного датасета**

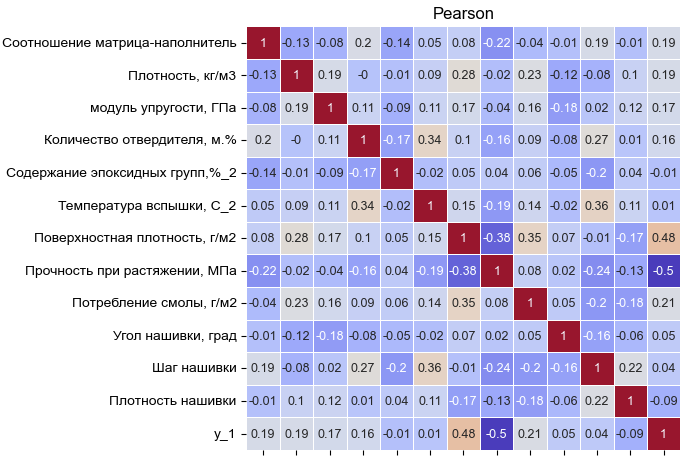


Рисунок 16 – Матрица корреляции признаков очищенного датасета (*knn* = 5)

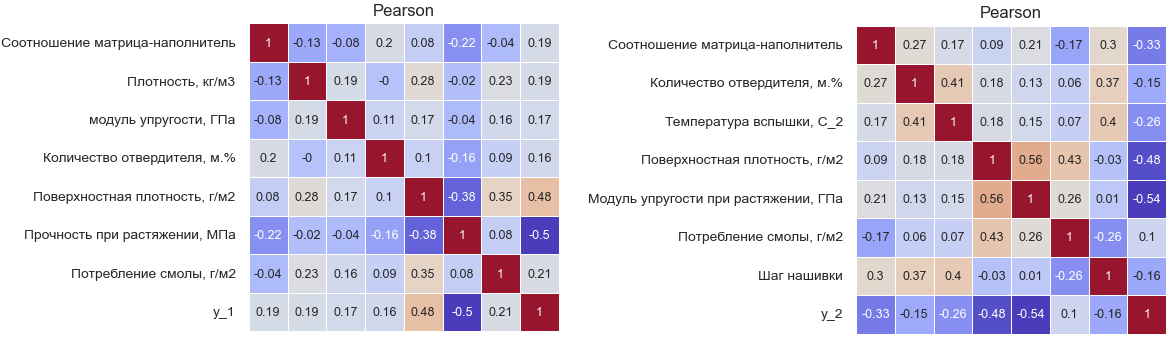


Рисунок 17 – Матрицы корреляции отобранных признаков для переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' (очищенные датасеты; *knn* = 5 и *knn* = 3, соответственно)

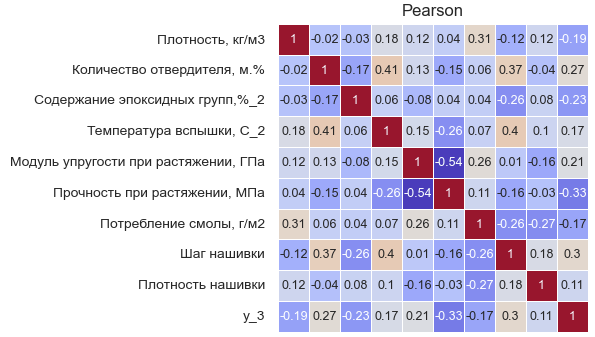


Рисунок 18 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (очищенный датасет; *knn* = 3)

Таблица 3 – Аналитические коэффициенты детерминации для целевых переменных (очищенный датасет)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Аналитический коэффициент детерминации для МНК *R*2 | | |
| до разбиения на train-  и test-выборки | train-выборка | test-выборка |
| 'Модуль упругости при растяжении' | 0,434 | 0,361 | 0,512 |
| 'Прочность при растяжении' | 0,523 | 0,397 | 0,812 |
| 'Соотношение матрица - наполнитель' | 0,265 | 0,242 | 0,482 |

**3 РЕЗУЛЬТАТЫ МАШИННОГО ОБУЧЕНИЯ МОДЕЛЕЙ**

Программный код и результаты машинного обучения моделей приведены в следующих файлах, расположенных на странице автора в репозитории GitHub – см. 4.2:

first\_23.ipynb

y\_1\_full.ipynb

y\_1\_filtered.ipynb

y\_2\_full.ipynb

y\_2\_filtered.ipynb

y\_3\_full\_NN.ipynb

y\_3\_filtered\_NN.ipynb

y\_3\_filtered\_Regressor.ipynb

**3.1 Результаты обучения моделей на данных базового датасета**

В результате машинного обучения модели LinearRegression получены выражения (3, 4) для линейных функциональных зависимостей целевых переменных 'Модуль упругости при растяжении, ГПа' и 'Прочность при растяжении, МПа' от признаков 'Поверхностная плотность, г/м2' и 'Потребление смолы, г/м2'. Полученные выражения подтверждают полученные ранее результаты расчета аналитического коэффициента детерминации *R*2 = 1 для МНК – см. 2.2 и таблицу 1.

В таблице 4 приведены метрики для обученных моделей различных регрессоров для переменной 'Соотношение матрица - наполнитель'. Максимальное значение коэффициента детерминации *R*2 = 0,505 для обучающей выборки (все значения базового датасета) имеет модель GradientBoostingRegressor.

Следует отметить, что значение коэффициента детерминации *R*2 = 0,393 для модели LinearRegression идентично значению аналитического коэффициента детерминации *R*2 для МНК – см. 2.2 и таблицу 1.

Доверительный интервал для расчета значений обучающей выборки переменной 'Соотношение матрица - наполнитель' по модели GradientBoostingRegressor приведен на рисунке 19. Все значения обучающей выборки расположены в пределах доверительного интервала, что позволяет сделать вывод о приемлемой точности обученной модели.

**3.1.1 'Модуль упругости при растяжении'**

y\_1 = 73,51 + 0,0128 ⸱ x\_6 – 0,0282 ⸱ x\_9 , (3)

где x\_6 = 'Поверхностная плотность, г/м2';

x\_9 = 'Потребление смолы, г/м2'.

**3.1.2 'Прочность при растяжении'**

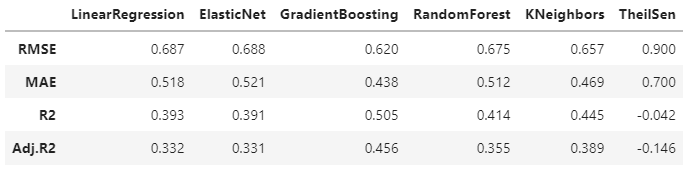
y\_2 = 1582,9 - 2,094 ⸱ x\_6 + 8,44 ⸱ x\_9 , (4)

где x\_6 = 'Поверхностная плотность, г/м2';

x\_9 = 'Потребление смолы, г/м2'.

**3.1.3 'Соотношение матрица - наполнитель'**

Таблица 4 – Метрики различных моделей для переменной 'Соотношение матрица - наполнитель' (базовый датасет)



****

Рисунок 19 – Доверительный интервал для расчета обучающих значений переменной 'Соотношение матрица - наполнитель' (базовый датасет; модель – GradientBoostingRegressor)

Примечание. Зеленым цветом показаны границы доверительного интервала

**3.2 Результаты обучения моделей на данных полного датасета**

В таблице 5 приведены метрики для различных обученных моделей для переменной 'Модуль упругости при растяжении'. Максимальное значение коэффициента детерминации *R*2 = 0,01 для тестовой выборки имеет модель KneighborsRegressor.

В таблице 6 приведены метрики для различных обученных моделей для переменной 'Прочность при растяжении'. Максимальное значение коэффициента детерминации *R*2 = 0,012 для тестовой выборки имеет модель TheilSenRegressor.

Следует отметить, что полученные значения коэффициентов детерминации *R*2 для обучающих значений модели LinearRegression идентичны соответствующим значениям аналитического коэффициента детерминации *R*2 для МНК – см. таблицу 2.

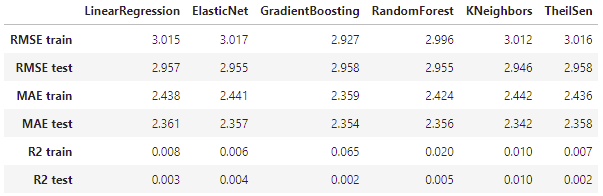
На рисунке 20 приведены метрики для обученной модели искусственной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для переменной 'Соотношение матрица - наполнитель'. Значение коэффициента детерминации для тестовой выборки составило *R*2 = 0,026.

Доверительный интервал для прогноза тестовых значений обучающей выборки переменной 'Соотношение матрица - наполнитель' приведен на рисунке 21. Практически все значения тестовой выборки расположены в пределах доверительного интервала.

Полученные метрики моделей, обученных на данных полного датасета, позволяют сделать вывод о том, что на этих данных **не удалось** построить нетривиальные (не константные), то есть пригодные для практического применения модели для целевых переменных. Данный вывод подтверждает, сделанный в 2.2 вывод о принципиальной невозможности или малой вероятности построения нетривиальных (не константных) моделей для целевых переменных по объектам полного датасета.

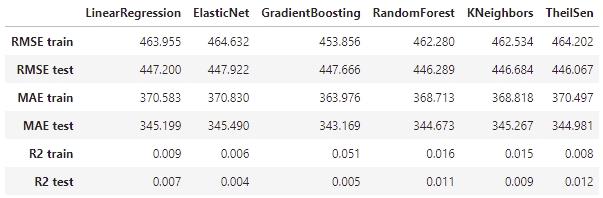
**3.2.1 'Модуль упругости при растяжении'**

Таблица 5 – Метрики различных моделей для переменной 'Модуль упругости при растяжении' (полный датасет)

****

**3.2.2 'Прочность при растяжении'**

Таблица 6 – Метрики различных моделей для переменной 'Прочность при растяжении' (полный датасет)

****

**3.2.3 'Соотношение матрица - наполнитель'**

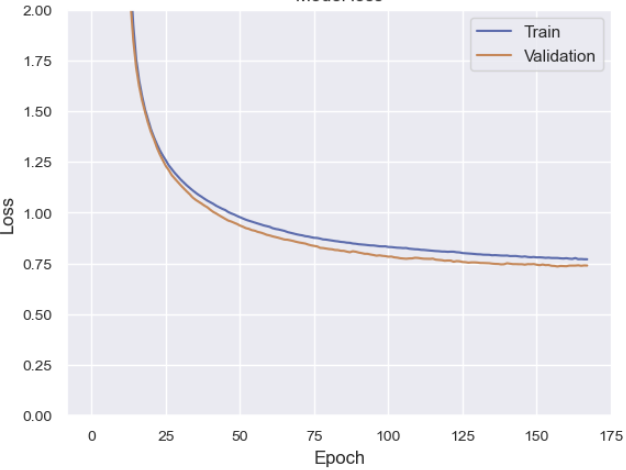
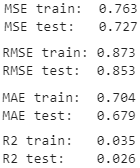
****

Рисунок 20 – Метрики и график функции потерь полносвязной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для значений переменной 'Соотношение матрица - наполнитель' (полный датасет)

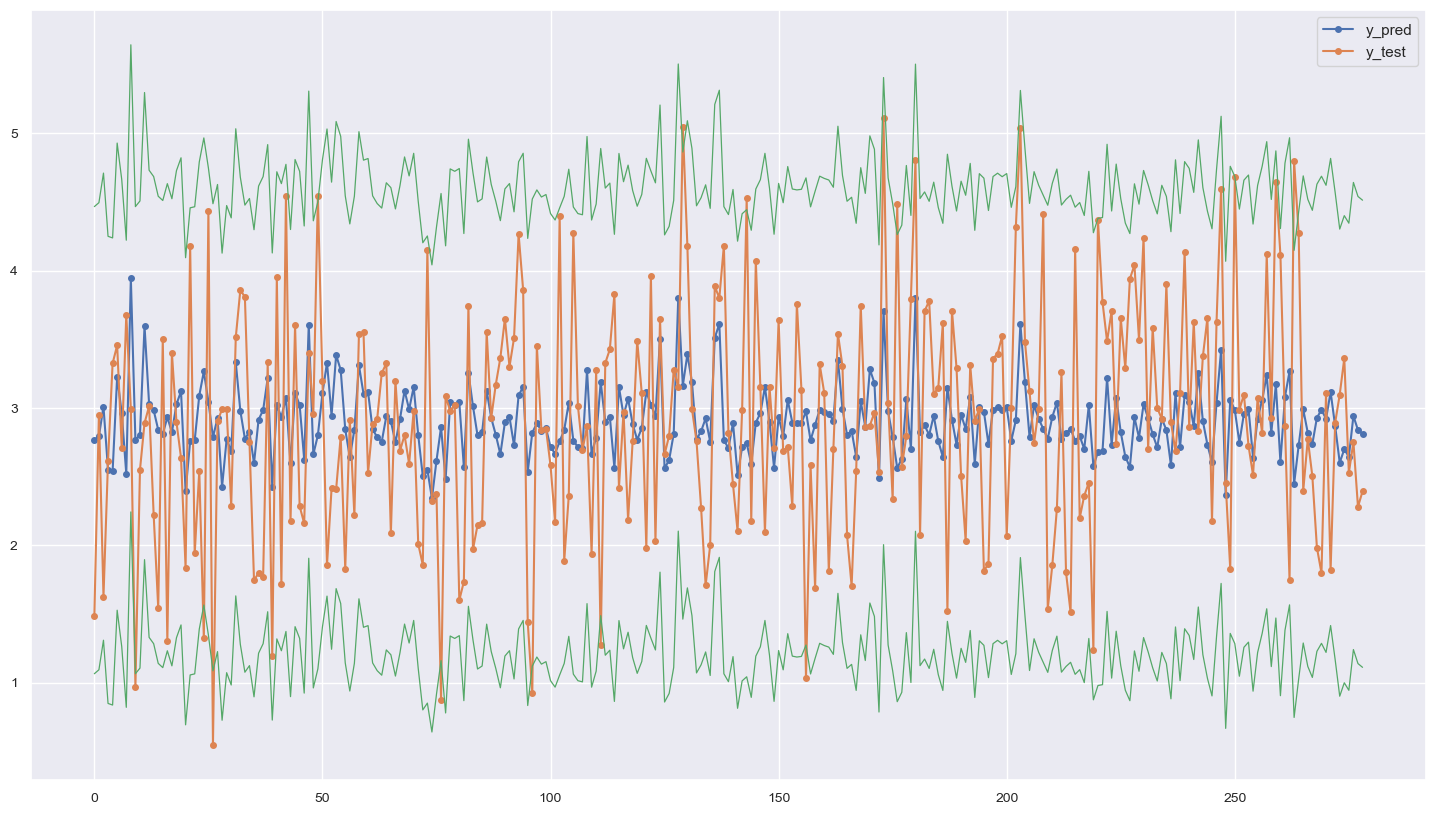
****

Рисунок 21 – Доверительный интервал для прогноза тестовых значений переменной 'Соотношение матрица - наполнитель' (полный датасет)

Примечание. Зеленым цветом показаны границы доверительного интервала

**3.3 Результаты обучения моделей на данных очищенного датасета**

В таблице 7 приведены метрики для различных обученных моделей для переменной 'Модуль упругости при растяжении'. Максимальное значение коэффициента детерминации *R*2 = 0,49 для тестовой выборки имеет модель GradientBoostingRegressor.

В таблице 8 приведены метрики для различных обученных моделей для переменной 'Прочность при растяжении'. Максимальное значение коэффициента детерминации *R*2 = 0,737 для тестовой выборки имеет модель RandomForestRegressor.

Следует отметить, что полученные значения коэффициентов детерминации *R*2 для обучающих значений модели LinearRegression идентичны соответствующим значениям аналитического коэффициента детерминации *R*2 для МНК – см. таблицу 3.

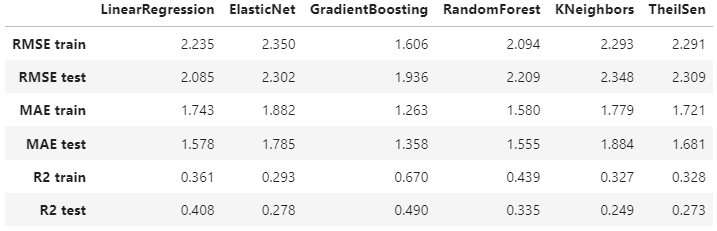
На рисунке 22 приведены метрики для обученной модели искусственной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для переменной 'Соотношение матрица - наполнитель'. Значение коэффициента детерминации для тестовой выборки составило *R*2 = 0,227.

Доверительный интервал для прогноза тестовых значений обучающей выборки переменной 'Соотношение матрица - наполнитель' приведен на рисунке 22. Все значения тестовой выборки расположены в пределах доверительного интервала, что позволяет сделать вывод о приемлемой точности обученной модели.

Полученные метрики моделей, обученных на данных очищенных датасетов, позволяют сделать вывод о том, что на этих данных **удалось** построить нетривиальные (не константные), то есть пригодные для практического применения модели для целевых переменных. Данный вывод подтверждает, сделанный в 2.2 вывод о принципиальной возможности построения нетривиальных (не константных) моделей для целевых переменных по объектам очищенных датасетов.

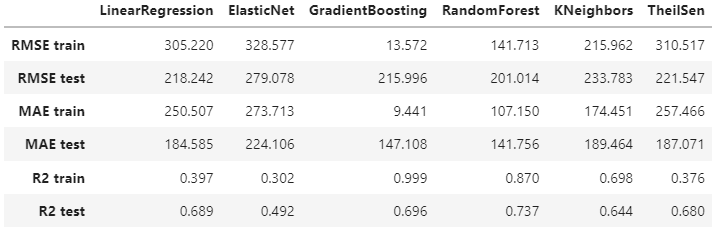
**3.3.1 'Модуль упругости при растяжении'**

Таблица 7 – Метрики различных моделей для переменной 'Модуль упругости при растяжении' (очищенный датасет: *N* = 108, *knn* = 5)

****

**3.3.2 'Прочность при растяжении'**

Таблица 8 – Метрики различных моделей для переменной 'Прочность при растяжении' (очищенный датасет: *N* = 76, *knn* = 3)

****

**3.3.3 'Соотношение матрица - наполнитель'**

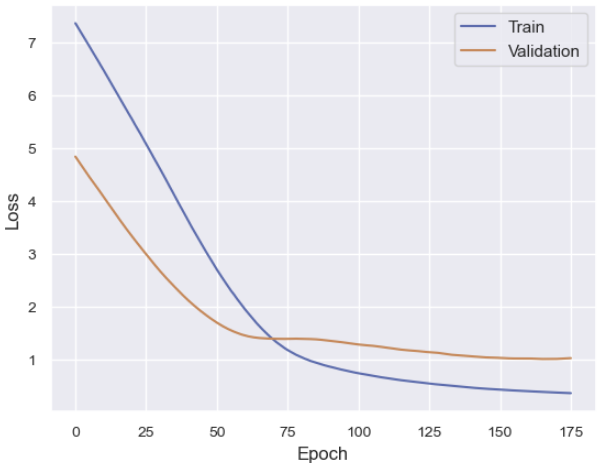
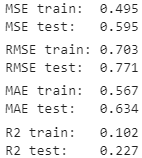
****

Рисунок 22 – Метрики и график функции потерь полносвязной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для значений переменной 'Соотношение матрица - наполнитель' (очищенный датасет)

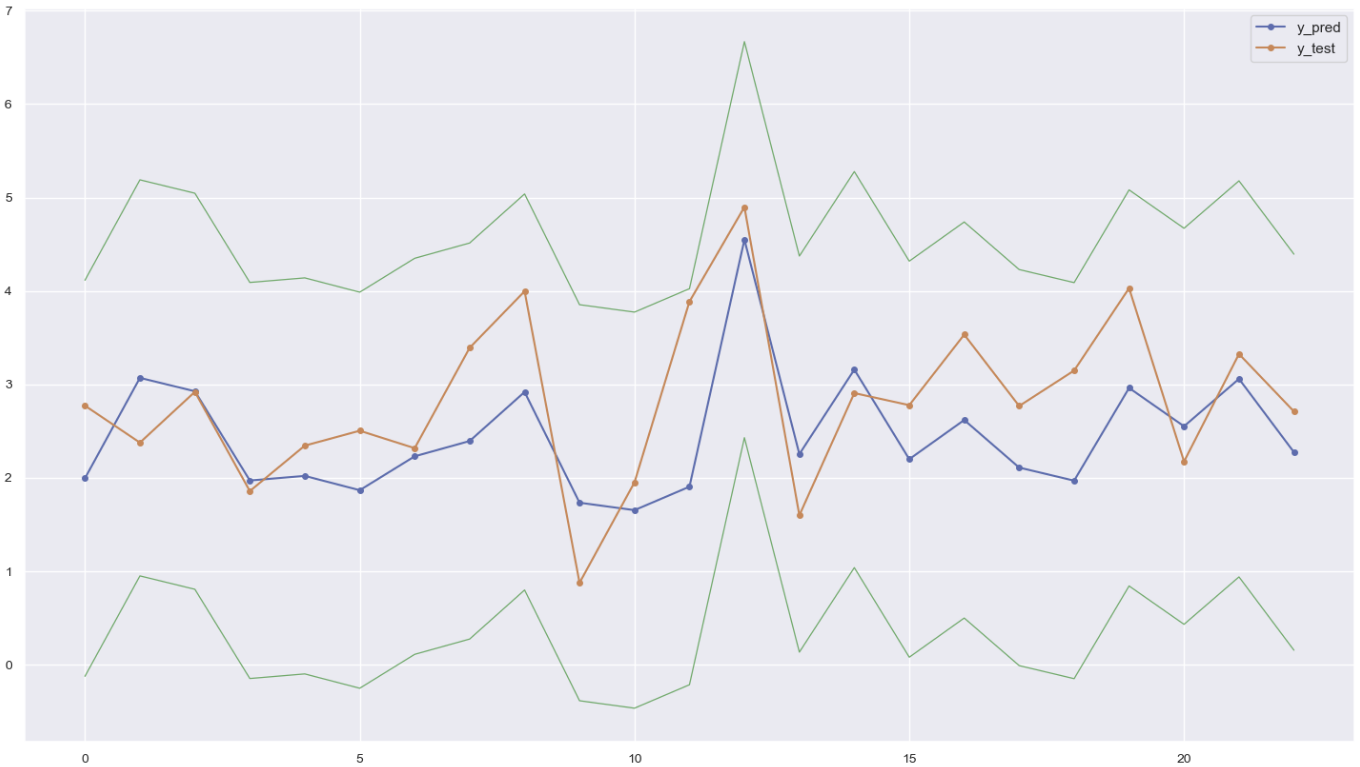
****

Рисунок 23 – Доверительный интервал для прогноза тестовых значений переменной 'Соотношение матрица - наполнитель' (очищенный датасет)

Примечание. Зеленым цветом показаны границы доверительного интервала

**4 СОЗДАНИЕ УДАЛЕННОГО РЕПОЗИТОРИЯ**

Результаты ВКР выложены на странице автора на сайте GitHub.com в репозитории BMSTU по адресу <https://github.com/GorshkovAndrey/BMSTU>

**4.1 Дата создания репозитория**

Репозиторий создан 08.04.2023.

**4.2 Состав репозитория**

Репозиторий состоит из трех папок:

1. Для защиты ВКР

В состав папки входят следующие файлы:

first\_23.ipynb

y\_1\_full.ipynb

y\_1\_filtered.ipynb

y\_2\_full.ipynb

y\_2\_filtered.ipynb

y\_3\_full\_NN.ipynb

y\_3\_filtered\_NN.ipynb

y\_3\_filtered\_Regressor.ipynb

Пояснительная записка ВКР. Горшков А.В..docx

Презентация ВКР. Горшков А.В..pptx

2. Первоначальные

В состав папки входят следующие файлы:

y\_1.ipynb

y\_2.ipynb

y\_3.ipynb

3. Фильтрация шума

В состав папки входят следующие файлы:

y\_1\_Lazy\_RdePRk5.ipynb

y\_2\_Lazy\_RdePRk3.ipynb

y\_3\_Lazy\_RdePRk3.ipynb

**ЗАКЛЮЧЕНИЕ**

1. Для построения математических моделей целевых переменных использованы три варианта набора данных – базовый датасет (первые 23 строки), полный датасет (1023 строки) и очищенный от шума и помех датасет.

2. Для очистки датасета от шума и помех разработан фильтр, ядром (маской) которого являются объекты базового датасета.

3. Для оценки принципиальной возможности построения нетривиальной (не константной) математической модели использовался аналитический коэффициент детерминации для МНК, вычисляемый по коэффициентам корреляции целевых переменных и признаков датасета.

В результате проведенных расчетов аналитического коэффициента детерминации сделан вывод о принципиальной невозможности или малой вероятности построения нетривиальных (не константных) моделей для целевых переменных по объектам полного датасета.

4. Для базового датасета получены линейные функциональные зависимости для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' от признаков 'Поверхностная плотность, г/м2' и 'Потребление смолы, г/м2'. Для целевой переменной 'Соотношение матрица - наполнитель' обучены модели различных регрессоров, для которых наибольший коэффициент детерминации составил = 0,505.

5. Для полного датасета для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' обучены модели регрессоров, для которых наибольший коэффициент детерминации составил = 0,012. Для целевой переменной 'Соотношение матрица - наполнитель' обучена модель искусственной нейронной сети, для которой коэффициент детерминации составил = 0,026.

Полученные метрики моделей, обученных на данных полного датасета, позволяют сделать вывод о том, что на этих данных **не удалось** построить нетривиальные (не константные), то есть пригодные для практического применения модели для целевых переменных. Данный вывод подтверждает приведенный выше вывод о принципиальной невозможности или малой вероятности построения нетривиальных (не константных) моделей для целевых переменных по объектам полного датасета.

6. Для очищенных датасетов для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' обучены модели различных регрессоров, для которых наибольший коэффициент детерминации составил  = 0,737. Для целевой переменной 'Соотношение матрица - наполнитель' обучена модель искусственной нейронной сети, для которой коэффициент детерминации составил = 0,227.

Полученные метрики моделей, обученных на данных очищенных датасетов, позволяют сделать вывод о том, что на этих данных **удалось** построить нетривиальные (не константные), то есть пригодные для практического применения модели для целевых переменных. Данный вывод подтверждает приведенный выше вывод о принципиальной возможности построения нетривиальных (не константных) моделей для целевых переменных по объектам очищенных датасетов.

**БИБЛИОГРАФИЧЕСКИЙ СПИСОК**

1. Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on.

2. Yeo, In-Kwon; Johnson, Richard A. (2000). "A New Family of Power Transformations to Improve Normality or Symmetry". [Biometrika](https://en.wikipedia.org/wiki/Biometrika" \o "Biometrika). 87 (4): 954–959.

3. Вентцель Е.С. Теория вероятностей: Учеб. для вузов. – 6-е изд. стер. – М.: Высш. шк., 1999

4. Theil. A rank-invariant method of linear and polynomial regression analysis. I, II, III // Nederl. Akad. Wetensch., Proc.. — 1950. —Т. 53. — С. 386–392, 521–525, 1397–1412.